SRI International

COMPUTATIONAL STEREO

Technical Note 261

March 1982

Stephen T. Barnard, Computer Scientist
Martin A. Fischler, Program Director, Vision

Artificial Intelligence Center
Computer Science and Technology Division

Support for the preparation of this paper was provided under DARPA Contract No. MDA903--79-C-0588.

SRI International®

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 1982		**00-03-1982 to 00-03-1982**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Computational Stereo	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**40**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

ABSTRACT

Perception of depth is a central problem in machine vision. Stereo is an attractive technique for depth perception because, compared to monocular techniques, it leads to more direct, unambiguous, and quantitative depth measurements, and unlike such "active" approaches as radar and laser ranging, it is suitable in almost all application domains.

We broadly define computational stereo as the recovery of the three-dimensional characteristics of a scene from multiple images taken from different points of view. The first part of the paper identifies and discusses each of the functional components of the computational stereo paradigm: image acquisition, camera modeling, feature acquisition, matching, depth determination, and interpolation. The second part discusses the criteria that are important for evaluating the effectiveness of various computational stereo techniques. The third part surveys a representative sampling of computational stereo research.

CONTENTS

ILLUSTRATIONS

I INTRODUCTION

This paper surveys and evaluates computational methods for the recovery of depth information from multiple images. We identify the major functional components that comprise these methods, list various alternative algorithms for implementing them, and discuss the domain-dependent and application-dependent constraints that favor some alternatives over others.

The scope of this paper is primarily restricted to research in the image understanding (IU) community. IU is a program of research in machine vision originated and largely supported by the Advanced Research Projects Agency (ARPA) of the Department of Defense. IU researchers have drawn on stereo work from other areas, especially cartography, psychology, and neurophysiology. We will not try to cover all the IU research relevant to stereo, but instead will select a cross-section of the most widely-known work that covers all the important and significantly different approaches to the stereo problem.

Much of the research in image-understanding has been devoted to recovering the range and orientation of surfaces and objects depicted in imaged data. The earliest work concentrated on an artificial domain -- the "blocks world" [Rob65]. Significant (but not necessarily extendable) advances were made in this simple domain; in particular, it was shown that edge and vertex labeling schemes could provide constraints that allowed one to correctly partition a complex scene. [Guz68], [Wal75]. More recent work, which has concentrated on real world problems, can be divided into three classes: (1) those methods that start with range information directly provided by an active sensor, (2) those methods that depend on monocular information available in a single image (or perhaps several images from a single viewpoint under different lighting), and (3) those methods that use two or more images taken from different viewpoints and perhaps at different times. We are concerned here with this third class, which we shall refer to as "generalized stereo."

The generalized stereo paradigm includes conventional stereo, as well as what is often called optic flow. In conventional stereo two images are recorded simultaneously by laterally displaced cameras (figure 1). In optic flow two or more images are recorded sequentially, usually with a single camera that moves along an arbitrary path. In a sense, conventional stereo can be considered to be a special case of optic flow, and the same geometrical formalisms apply to both.

Stereo is an attractive source of information for machine perception becase it leads to direct range measurements, and, unlike monocular approaches, does not merely infer depth or orientation through the use of photometric and statistical assumptions. Once the stereo images are brought into point-to-point correspondence, the recovery of range values is a relatively straightforward matter. Furthermore, stereo is a passive method. Active ranging methods that use structured light, laser rangefinders, or other active sensing techniques are useful in tightly controlled domains, such as industrial automation applications, but are clearly unsuitable for more general machine vision problems.

Perhaps the most common use of computational stereo is in the interpretation of aerial images. Other applications are passive navigation for autonomous vehicle guidance, and industrial automation applications. Each domain has different requirements that can affect the design of a complete stereo system.

II THE COMPUTATIONAL STEREO PARADIGM

Research on computational solutions for the generalized stereo problem has followed a single paradigm, although there have been several distinct variations, both in method and intent. The paradigm involves the following steps.

* Image acquisition
* Camera modeling

* Feature acquisition
* Matching
* Distance (depth) determination
* Interpolation

A. Image Acquisition

The most important factor affecting image acquisition is the specific application for which the stereo computation is intended. Three applications have received the most attention: the interpretation of aerial photographs for automated cartography; guidance and obstacle avoidance for autonomous vehicle control; and the modeling of human stereo vision.

Aerial photo-interpretation usually involves low-resolution images of a variety of terrain types. Aerial stereo images may be either vertical, in which the camera axes point vertically downward as nearly as possible, or oblique, in which the camara axes are intentionally directed between the horizontal and vertical directions (figure 2). Vertical stereo images are easier to compile into precise cartographic measurements, but oblique stereo images cover more terrain and require less stringent control of the aircraft.

Stereo for autonomous vehicle control has been studied in two contexts: as a passive navigation aid for drone aircraft [Han80], and as part of a control system for surface vehicles [Mor79], [Mor81], [Genn80]. The images used for aircraft navigation are similar to the aerial photographs used for cartography, except that long sequences of images are used, and multispectral sensors are often employed. The images used for surface vehicle control are quite different -- they are horizontal, comparatively high-resolution images.

Research on computational models of human stereo vision has largely employed synthetic random-dot stereograms for experimental investigation [Marr76], [Marr77], [Grim79], [Grim80], [Grim81]; the primary reason for this is that random-dot stereograms exclude all monocular depth cues, and the exact correspondences are known. Because the parameters of

random-dot stereograms, such as noise and density, can be controlled, they allow systematic comparison of human and machine performance. This does not imply that experiments with natural imagery has been ignored in research on human stereo vision (for example, see [Grim80]).

Perhaps the most significant and widely recognized difference in scene domains is the difference between scenes containing cultural features such as buildings and roads, and those containing only natural objects and surfaces such as mountains, flat or "rolling" terrain, foliage, and water. Important stereo applications range over both domains. Low-resolution aerial imagery, for example, usually contains mostly natural features, although cultural features are sometimes found. Industrial applications, on the other hand, tend to involve man-made objects exclusively. Cultural features present special problems. For example, periodic structures such as the windows of buildings and road grids can confuse a stereo matcher. The relative abundance of occlusion edges in a city scene also causes problems because large portions of the images may be unmatchable. Cultural objects often have large surfaces with nearly uniform albedo that are difficult to match because of a lack of detail. Stereo systems that have been described in the literature are usually targeted at specific scene domains, and there is seldom any attempt to validate the methods in other domains.

In summary, the key parameters associated with image acquisition are:

* Scene domain
* Timing
  - Simultaneous
  - Nearly simultaneous
  - Radically different times
* Time of day (lighting and presence of shadows)
* Photometry (including spectral coverage)
* Resolution
* Field of view
* Relative camera positioning (length and orientation, relative to the scene, of the stereo base line).

The issues associated with the scene domain are percentage of:

* Occlusion
* Man-made objects (straight edges, flat surfaces)
* Continuous surfaces of some minimal extent
* Textureless area
* Area containing repetitive structure.

B. Camera Modeling

The key problem in automated stereo is to find corresponding points in the stereo images. Corresponding points are the projections of a single point in the three-dimensional scene. The difference in the positions of two corresponding points in their respective images is called "parallax" or "disparity". Disparity is a function of both the position of the point in the scene, and of the position, orientation, and physical characteristics of the stereo cameras. When these camera attributes are known, corresponding image points can be mapped into three-dimensional scene locations. A camera model is a representation of the important geometrical and physical attributes of the stereo cameras. It may have a relative component, which relates the coordinate system of one camera to the other, and is independent of the scene; and it may have an absolute component, which relates one of the camera coordinate systems to the fixed coordinate system of the scene.

In addition to providing the function which maps pairs of correponding points into scene points, a camera model can be used to constrain the search for matches of corresponding points to one dimension (figure 3). Any point in the three-dimensional world space, together with the centers of projection of two camera systems, defines an epipolar plane. The intersection of an epipolar plane with an image plane is called an epipolar line. Every point on a given epipolar line in one image must correspond to a point on the corresponding epipolar line in the other image. The search for a match of a point in the first image may therefore be limited to a one-dimensional neighborhood in the second image plane, as opposed to a two-dimensional neighborhood, with an enormous reduction in computational complexity.

When the stereo cameras are located and oriented such that there is only a horizontal displacement between them, then disparity can only occur in the horizontal direction, and the stereo images are said to be "in correspondence." When a stereo pair is in correspondence the epipolar lines are coincident with the horizontal scan lines -- a convenient situation because the matching process can be accomplished in a relatively simple and efficient manner. Stereo systems that have been primarily concerned with modeling human ability have employed this constraint [Grim80,Marr77]. In practical applications, however, the stereo pair may not be in correspondence. In aerial stereo photogrammetry, for example, the camera axis may typically be tilted as much as two to three degrees from vertical [Thom66]. The implication here is that points on a scan line in one image will not fall on a single scan line in the second image of the stereo pair, and thus, the computational cost to employ the epipolar constraint is significantly increased. It is possible, however, to reproject the stereo images onto a common plane parallel to the stereo baseline such that they are in correspondence.

The difference in position and orientation of two stereo cameras is called the relative camera model. Relative camera models are required for depth determination, and also allow one to exploit the epipolar constraint. In most cases, considerable *a priori* knowledge of the relative camera model is available, but it is often not as accurate as desired. Gennery [Genn79] has developed a method for solving for the relative camera model from a few sparse matches. His method accounts for differences in azimuth, elevation, pan, tilt, roll, and focal length (figure 4).

Fischler and Bolles [Fisc81] have provided a number of results with respect to the minimum number of points needed to obtain a solution to the camera modeling problem, given a single image and a set of correspondences between points in the image and their spatial (geographic) locations; they also provide a technique for solving for the complete camera model, even when the given correspondences contain a

large percentage of errors. While this work was directed at the problem of establishing a mapping between an image and an existing geographic database, it is possible to apply the results to the stereo problem, and in fact, tying the stereo pair to an existing database offers the possibility of employing scene dependent constraints beyond those available from the imaging geometry.

Camera modeling can be extended to include distortions introduced in the image-making process. Significant image distortion will degrade the accuracy of depth measurements made by a stereo system unless corrected. Two kinds of image distortion are commonly found: radial and tangential. Radial distortion causes image points to be displaced perpendicular to the optical axis and may occur in the form of pin-cushion distortion (i.e., positive radial distortion) or barrel distortion (i.e., negative radial distortion). Tangential distortion is caused by imperfect centering of lens elements, resulting in image displacements perpendicular to the radial lines. Moravec described a method to correct for distortion using a square pattern of dots [Mor79]. Fourth degree polynomials are found that transform the measured positions of the dots and their neighborhoods to their nominal positions.

In summary, the important issues in camera modeling are:

* A priori knowledge of camera positions and parameters
* Solutions using a few sparse matches
* A priori knowledge of the geographic locations (three-dimensional scene coordinates) of selected scene objects and features
* Ability to deal with matching errors
* Compensation for image distortion

C. Feature Acquisition

Featureless areas of nearly homogeneous brightness cannot be matched with confidence. Accordingly, most work in computational stereo has included some form of selective feature detection, the particular form of which is closely coupled with the matching strategy used.

Approaches that apply area matching often use an "interest operator" to locate places in one image that can be matched with confidence to corresponding points in the second image of a stereo pair. One way to do this is to select areas that have high image intensity variance. These areas will not be good features, however, if the variance is due only to brightness differences in the direction perpendicular to the epipolar line. These areas can be culled by demanding that the two-dimensional autocorrelation function have a distinct peak [Han74]. A widely used interest operator is the Moravec operator [Mor79], which selects points that have high variance between adjacent pixels in four directions. Hannah has modified this operator to consider ratios of the variances in the four directions, as well as ordinary image intensity variance over larger areas, and this modified operator seems to locate a better selection of both strong and subtle features [Han80].

Feature detection is more centrally important to those approaches that directly match features in the stereo images (rather than simply using the features to choose areas for correlation matching). The features may vary in size, direction, and dimensionality. Point-like features are good candidates for matching when the camera model is unknown and the matches are not constrained to epipolar lines. This is because, unlike linear features, points are unambiguously located in the image and can be matched in any direction. Linear features must be oriented across the epipolar lines if they are to be matched accurately. An advantage of point-like features is that they can be matched without concern for perspective distortion. In area-correlation approaches point-like features are often used to obtain the camera model prior to more extensive matching. The local intensity values around a point can

be used to establish initial confidences of matches in a way similar to area correlation [Bar80].

If the camera model is known a priori or derived in a preliminary step, edge elements can be used as primitive matching features. Many distinct edge models have been proposed as the basis for edge-detecting algorithms. In the case of "strong" edges, most of the resulting algorithms yield similar results for operators of comparable sizes. Often the same underlying model appears in different implementations; e.g., zero-crossings in the second derivative are equivalent to local maxima in the first derivative, and most of the conventional edge detection methods search for approximations to maxima of the first derivative of image intensity. More important are the issues governing the conditions under which "weak" edges found by different algorithms are reliable features for matching. Size, direction, and magnitude (i.e., contrast) have been been used as features in making match decisions, but their relative merit is not established.

For the most part, low level features have been used for stereo. What we mean by "low level" is that the features depend only on local monocular intensity patterns, and are based on the assumption that more-or-less sharp intensity gradients are due to physically significant structural, reflectance, and illumination events in the scene (as opposed to being artifacts of the camera location). Higher level features that depend on more sophisticated semantic analysis have been largely unused (Ganaparthy described a system for matching vertices in blocks-world stereo scenes across very large viewing angles [Gan75]). The ability to classify edges as occlusion or nonocclusion boundaries [Wit81], for example, could be very useful to a stereo system, especially in the difficult domains that include a wealth of cultural features.

In summary, the properties of local features that are important to the computational stereo problem are:

* Dimensionality (point-like versus edge-like)
* Size (spatial frequency)
* Contrast

* Semantic content
* Density of occurence
* Easily measurable attributes
* Uniqueness/distinguishability.

D. Matching

Image matching is a core area in scene analysis and will not be covered in in full detail in this paper. Instead, we will focus on those portions of the image-matching problem that are directly relevant to stereo modeling. Features that distinguish stereo image matching from image matching in general are the following:

* The important differences in the stereo images are due to the different viewpoints, and not, for example, due to changes in the scene. We therefore seek a match between two images, as opposed to a match between an image and an abstract model (although matching to an abstract model may be an important step in determining the image-to-image matching).
* Most of the significant changes will occur in the appearance of nearby objects and in occlusions. Additional changes in both geometry and photometry can be introduced in the film development and scanning steps, but can usually be avoided by careful processing. If the images are recorded at very different times there may be significant lighting effects.
* Stereo modeling generally requires that, ultimately, a dense grid of points be matched.

Ideally, we would like to find the correspondences (i.e., the matched locations) of every individual pixel in both images of a stereo pair. However, it is obvious that the information content in the intensity value of a single pixel is too low for unambiguous matching. In practice, coherent collections of pixels are matched. These collections are determined and matched in two distinct ways:

* Area Matching: Regularly sized neighborhoods of a pixel are the basic units that are matched. This approach is justified by the "continuity assumption," which asserts that at the level of resolution at which stereo matching is feasible, most of the image depicts portions of continuous surfaces; therefore, adjacent pixels in an image will

generally represent contiguous points in space. This approach is almost invariably accompanied by correlation based matching techniques to establish the correspondences.

* Feature Matching: "Semantic features" (with known physical properties and/or spatial geometry), or "intensity anomaly features" (isolated anomalous intensity patterns not necessarily having any physical significance), are the basic units that are matched. (See the discussion in the preceding section on feature acquisition.) Semantic features of the generic type include occlusion edges, vertices of linear structures, and prominent surface markings; domain-specific semantic features might include, for example, the corner or peak of a building, or a road surface marking; intensity anomaly features include zero-crossings and image patches found by the Moravec interest operator. Methods used for feature matching often include symbolic classification techniques, as well as correlation.

Obviously, feature matching alone cannot provide the desired dense depth map so it must be augmented by a model-based interpretation step (e.g., we recognize the edges of buildings and assume that the intermediate space is occupied by planer walls and roofs), or by area matching. When used in conjunction with area matching, the feature matches are generally considered to be more reliable and can constrain the search for correlation matches.

To further reduce the possibility of error caused by an ambiguous match, a number of hierarchical and global matching techniques have been employed, including relaxation matching and various "coarse-fine" hierarchical matching strategies.

The correlation-matching approach attempts to resolve ambiguity by using as much local information as possible to make decisions about potential matches, but each match decision is made independently of the others. The relaxation-labeling approach [Bar80] uses a relatively small amount of local information for each potential match, and attempts to resolve ambiguity by finding consensuses among subsets of the total population of matches. It relies on the three-dimensional continuity of surfaces to be reflected in the two-dimensional continuity of disparity. A method for avoiding ambiguity that can be applied to both correlation

matching [Mor79] and feature point matching [Marr77] is the so-called "coarse-fine" strategy. In this approach coarse disparities are found relatively quickly, but with low accuracy. These gross disparities are used to constrain finer-resolution matching. Even with a coarse-fine strategy, however, some ambiguity at each level of resolution is inevitable. The best combination of ambiguity avoidance and ambiguity resolution is a major research issue.

Matching is complicated by several factors related to the geometry of the stereo images. Some areas that are visible in one image may be occluded in the other, and this can lead to incorrect matches. Periodic structures in the scene can confuse a matcher when the image features generated by these structures are close together compared to the disparity of the features, because the matcher may confuse a feature in one image with features from nearby parts of the structure in the other image. If there is a large amount of relief in the scene (for example, a vertical obstruction that projects above the ground plane in an aerial view) then corresponding features may actually be reversed in their positions in the two stereo images.

In summary, key attributes which differentiate matching techniques include:

* Local versus global ambiguity resolution
* Area (dense) versus feature (sparse) matching.

The constraints used to both limit computation and reduce ambiguity include:

* Epipolar
* Continuity
* Hierarchical (e.g., coarse-fine matching)
* Sequential (e.g., feature tracking in sequential views).

Criteria that can be used to evaluate (or compare) different matching techniques include:

* Accuracy (match precision measured to the sub-pixel level)
* Reliability (resistance to gross classification errors)
* Generality (applicability to different scene domains)

* Predictability (availability of performance models)
* Complexity (cost of implementation; computational requirements).

E. Distance Determination

With few exceptions, work in image understanding has not dealt with the specific problem of distance determination. The matching problem has been considered the hardest and most significant problem in computational stereo. Once accurate matches have been found the determination of distance is a relatively simple matter of triangulation. Nevertheless, this step presents significant difficulties, especially if the matches are somewhat inaccurate or unreliable.

To a first approximation, the error in stereo distance measurements is directly proportional to the positional error of the matches and inversely proportional to the length of the stereo baseline. Lengthening the stereo baseline complicates the matching problem by increasing the range of disparity (i.e., the area that must be searched) and the difference in appearance of the features being matched. Various matching strategies have been used to overcome this problem (coarse/fine strategies, cooperative or relaxation-labeling approaches, and matching of several incremental stereo views).

In many cases, matches are made to an accuracy of only a pixel. However, both the area correlation and the feature-matching approaches can provide better accuracy. Sub-pixel accuracy using area correlation requires interpolation over the correlation surface. Some feature detection methods can locate features to accuracies better than one pixel, but this depends on the type of operator that is used, and there are no generally applicable techniques.

Another approach is to settle for one-pixel accuracy, but to use multiple views [Mor79]. A match from a particular pair of views represents a depth estimate with uncertainty that depends on the accuracy of the match and on the length of the stereo baseline. Matches

from many pairs of views can be statistically averaged to find a more accurate estimate. The contribution of a match to the final depth estimate can be weighted according to any of the factors that bear on the confidence of the match and on its accuracy.

In summary, improved depth measurements can be obtained in several ways, each involving some additional computational cost:

* Sub-pixel estimation
* Increased stereo baseline
* Statistical averaging over several views.

F. Interpolation

As previously mentioned, stereo applications usually demand a dense array of depth estimates that the feature matching approach cannot provide because features are sparsely and irregularly distributed over the images. The area correlation-matching approach is more suited to obtaining dense matches, although it tends to be unreliable in areas of low information. Consequently, some kind of interpolation step is usually required.

The most straightforward way to create the dense depth array from a sparse array is simply to treat the sparse array as a sampling of a continuous depth function, and to approximate the continuous function using a conventional interpolation method (for example, by fitting splines). Assuming the sparse depth array is complete enough to capture the important changes in depth, this approach may be adequate. Aerial stereophotographs of rolling terrain, for example, might be handled in this way. In many applications, however, the continuous depth function model will not be appropriate because of occlusion edges.

Grimson [Grim81] has noted that the absence of matchable features implies a limit on the variability of the surface to be interpolated, and has proposed an interpolation procedure based on this observation. From a slightly different point of view, monocular "shape-from-shading" techniques (e.g., [Horn75]), employing the matched features to establish

boundary conditions, and the smooth intervening surface to assure the validity of integration, can provide an interpolation procedure with an acceptable physical justification.

Another approach to the interpolation problem is to fit a priori geometric models to the sparse depth array. Normally, model fitting would be preceded by clustering to find the subsets of points in the three-dimensional world space that correspond to significant structures in the scene. Each cluster would then be fit to the best available model, thereby instantiating the model's free variables and providing an interpolation function. This approach has been used to find ground planes [Arn78], elliptical structures in stereophotographs [Gen80], and smooth surfaces in range data acquired with a laser rangefinder [Duda79].

III EVALUATION CRITERIA

In evaluating the effectiveness of various computer stereo techniques we must consider a wide range of performance metrics. We must consider both quantitative measurements, such as accuracy, as well as fundamentally qualitative but nonetheless important measurements, such as domain sensitivity. Finding an optimum combination of techniques for an integrated system is difficult because of complex trade-offs in a large design space. The following criteria are appropriate for evaluating both complete stereo systems and the components of such systems. More specialized criteria relevant to individual components of stereo systems were presented in previous sections of this paper.

(1) Disparity - what range of disparity is handled? One possible advantage of automated stereo analysis is that computer methods may be able to handle larger angular disparities than humans can. Larger disparities lead to more accurate depth measurements, but also to more difficult matching problems.

(2) Coverage - what percentage of the scene is matched? Also, how widely are the matches distributed? Clearly,

large, featureless, homogeneous areas cannot be readily matched. What kinds of interpolation techniques can be used in such areas? What monocular techniques can be used to enhance coverage (for example, photometric evidence for smooth surfaces)?

(3) Accuracy

(4) Reliability - how many false matches are made compared to valid matches? What methods are effective for detecting and eliminating false matches?

(5) Domain sensitivity - what range of scene domains can be handled?

(6) Efficiency - actual timings of stereo systems will probably not be useful because of nonoptimal implementations and differences in hardware. Comparisons based on computational complexity can be made, however. How does the time required for stereo compilation scale with the image size, with the range of disparity, and with other important parameters? How amenable to hardware implementation are the different methods? What efficiency is needed for useful automated stereo systems?

(7) Human engineering - how are the results displayed (perspective 3D plots, false coloring, countour plots, vector fields, etc.)? What are the best methods? Is human interaction allowed?

(8) Sources of data for experimental validation - what kind of three-dimensional measurements are used to test performance? Three possibilities are:

Synthetic images or images of scaled models.
* Advantages: cheap, certainty about actual depths, control over secondary parameters
* Disadvantage: not representative of any real image domain

Ground surveys.
* Advantages: realistic, certainty about actual depths
* Disadvantage: expensive (hence limited number of sites that can be surveyed)

Compare to human performance.
* Advantages: realistic, reasonably inexpensive
* Disadvantages: susceptible to human errors, limited accuracy

IV SURVEY

This survey covers a representive sampling of the image understanding work relevant to computational stereo. While not exhaustively covering the field, it does contain examples of all the significantly different approaches to the steps in the computational stereo paradigm. The work discussed in the survey is grouped according to the research centers where the primary investigators were resident.

Carnegie-Mellon University

An iterative image registration technique with an application to stereo vision, [Luc81]

The emphasis in this work is on image registration, but there is also direct application to stereo matching. The general approach is to refine an estimate of the disparity of a region by using image intensity gradient information. This is done by inferring a correction to the a priori disparity of the region from the local intensity differences between the images and from the intensity gradient of one of the images. The correction is computed iteratively until the disparity converges to a final estimate. This method is closely related to a class of image matching techniques introduced by Limb and Murphy [Limb75]. A similar technique was used by Fennema and Thompson [Fenn79] to match images of moving objects. The method can be used to find not only disparity, but also brightness and contrast differences between the images, as well as the parameters relating the two camera systems (in conjunction with the relative camera model solution presented in [Genn79]).

The algorithm will converge to the correct answer when the disparity is no larger than one-half of the wavelength of the largest frequency component in the images. This implies that the method should be used with a coarse-fine strategy. It will not work well where there are sharp changes in depth, such as at the edge of an object.

Control Data Corporation

A flexible approach to digital stereo mapping, [Henn78]

This work is concerned with the automation of stereo-mapping functions. The primary concerns have been with handling different kinds of terrain and sensors, efficient hardware implementation, and the development of an interactive mapping system.

A regularly spaced grid of points in the left image is matched in the right image. Matching is accomplished by searching along the corresponding epipolar line in the right image for a maximum correlation patch, which is warped to account for predicted terrain relief (estimated from previous matches). Sub-pixel matches are obtained by fitting a quadratic to the correlation coefficients and picking the interpolated maximum.

"Tuning parameters" may be dynamically altered to adapt the system to sensor and terrain variations. Tuning parameters include grid sizes; patch size and shape; number of correlation sites along the search segment; and reliability thresholds for the correlation coefficient, standard deviation, prediction function range, and slope of the correlation function. The intent is to choose the smallest feasible patch, subject to the need to compensate for noise and lack of intensity variation in the image.

A continuity constraint is used to limit the search for matches. The rate of change of disparity is assumed to be continuous. This constraint is also used to shape the correlation patches in the left image.

The reliability of matching is continuously monitored to signal when parameters become inappropriate or when the photometry prevents valid matching. Reliability is estimated with a combination of factors, including correlation coefficients, patch standard deviation (are features present?), distances of actual from predicted correlation maxima, and slopes of the correlation functions.

The system is implemented on a highly parallel configuaration of 4 CDC Flexible Processors, each capable of 8 MIPS.

A somewhat different approach has been taken for three-dimensional modeling of cultural sites (e.g., building complexes) from high-resolution images. The basic idea is to identify corresponding points of intersection between epipolar lines and edges in the two images of a stereo pair. Non-matched edges are assumed to be due to noise or occlusions. Depth along an epipolar line (corresponding to a three-dimensional profile line in the scene) is assumed to vary linearly between contiguous pairs of matched intersections. Special techniques are developed to deal with occlusions and "reversals." Edge-tracking across sequential epipolar lines (the continuity constraint) contributes to reliability.

Lockheed

Bootstrap stereo, [Han80]

The goal of this study is navigation of an autonomous aerial vehicle using passively sensed images. Ground control points are used to determine the vehicle's initial location, and the corresponding camera model is used to locate further control points. The process can be iterated to continuously find new control points along the flight path. Major components of the system are camera calibration, new control point selection, matching, and control point position determination. The complete system consists of several navigation "specialists," including ones using instrumentation (altimeter, airspeed indicator, attitude gyros), dead reckoning, landmarks, and stereo.

Camera calibration is achieved with standard least-squares methods to determine position and orientation of the camera.

New control point selection involves an adaptation of the Moravec operator that uses ratios of variance along pairs of orthogonal directions (instead of simply the variance in four directions).

Control point matching is accomplished with normalized cross-correlation using a spiraling grid search. Coarse matching is used to approximately register the images and to initialize second-order prediction polynomials. Autocorrelation in the neighborhood of a matched point is used to evaluate the match. (The autocorrelation score indicates what constitutes a "good" match, and can therefore be used to select matches. This is a better alternative than using a global threshold on the normalized cross-correlation score.) Subpixel matching accuracy is achieved through parabolic interpolation of the correlation values.

The University of Minnesota

The Image Correspondence Problem, [Barr79]
Disparity analysis of images, [Bar80]

Points are matched in two images that differ because of stereo or object motion. The Moravec operator is used to select point features in both images. An initial collection of possible matches is established by linking each point in the first image with possible matching points in the second image. (A point in the second image is considered a possible match if it is in a square area centered on the position of the point in the first image.) Each point from the first image is considered an object that is to be classified according to its disparity, and each of its possible matches establishes a label denoting one of several possible classifications. Each object also has a special label denoting "no-match." An initial confidence for each disparity label is determined based on the mean-square-difference of small regions surrounding the possible matching points. The estimates are iteratively improved with a relaxation-labeling algorithm that uses the continuity constraint. Support for each label of a particular object is calculated from the neighboring objects. If relatively many nearby objects have similar labels with high confidence, the label is strongly supported and its confidence increases. If no labels are strongly supported, the

confidence of the "no-match" label increases. After a few iterations (about 8) the confidence estimates converge to unique disparity classifications for each point. (Convergence is not guaranteed theoretically, but is observed experimentally).

MIT

(1) Cooperative computation of stereo disparity, [Marr76]

A parallel, "cooperative" computational model for human stereo vision is proposed. This feature matching method uses two constraints to match random-dot stereograms. The features that are matched are the dots themselves. The constraints are: uniqueness, which requires that every feature have a unique disparity (a consequence of imaged points on three dimensional surfaces having unique depths); and continuity, which requires that disparity varies smoothly almost everywhere (except at relatively rare occlusion boundaries). These constraints are applied locally over several iterations with an algorithm very much like relaxation-labeling. Multiple disparity assignments of a point inhibit one-another, and local collections of similar disparities support one-another. Although this algorithm successfully fused random-dot stereograms, the authors rejected it as a model of human stereopsis and proposed a new model described below.

(2) A computational theory of human stereo vision, [Marr77]
A computer implementation of a theory of human stereo vision, [Grim79]
Aspects of a theory of human stereo vision, [Grim80]
From Images to Surfaces, [Grim81]

Matching of features occurs at different spatial scales. The matches found at the larger scales establish a rough correspondence for the smaller scales, thereby reducing the number of false matches. Features of different sizes are found by convolving the image with the Laplacian of a Gaussian mask. The size of the features is determined by the standard deviation (the "size constant") of the Gaussian mask.

Masks of four different sizes are used, separated by one spatial octave (i.e., each mask is twice the size of the next smaller one). Features are selected at the zero-crossings in the result (i.e., where the result changes sign). The zero-crossings after a second difference operation correspond to extrema after a first difference operation. This method is therefore a way of finding edge-like features at different scales. In the implementation a true Laplacian operator was not used; instead, the difference of two circularly symmetric Gaussians was used as a close approximation. The convolutions were done on a LISP machine augmented by special-purpose hardware. In the original theory, line terminations were to be used as features, along with zero-crossings, but this has not been implemented.

Zero-crossings where the gradient is oriented vertically are ignored (The implicit camera model has the epipolar lines oriented horizontally.). Other zero-crossings are located to an accuracy of one pixel and their orientations (determined by the gradient of the convolution values) are recorded in increments of 30 degrees. It is possible to interpolate the location of a zero-crossing to better than one pixel accuracy.

Matching within at any given scale proceeds independently of other scales. First, a zero-crossing is located in one image. The region surrounding the same location in the second image is then divided into three pools -- two larger "convergent" and "divergent" pools, and a smaller zero-disparity pool centered on the predicted match location. The three pools together span a region twice the width of the central positive region of the convolution mask. Zero-crossings from pools in the second image can match the zero-crossing from the first image only if they result from convolutions of the same size mask, have the same sign, and have approximately the same orientation. If a unique match is found (i.e., only one of the pools has a zero-crossing satisfying the above criteria), the match is accepted as valid. If two or three candidate matches are found, they are saved for future disambiguation. Once all matches have been found (ambiguous or not), the ambiguous ones

are resolved by searching through the neighborhoods of points to determine the dominant disparity (convergent, divergent, or zero). This is the familiar continuity constraint.

It may be the case that the disparity of a region is greater than the range handled by the matcher. This is detected from the percentage of unmatched zero-crossings. Marr and Poggio showed that the probability of a zero-crossing having at least one candidate match if the disparity is outside the range of the matcher is about 0.7, but is much higher if the disparity is within the range of the matcher.

The lower frequency matching channels are used to bring the higher frequency channels into range. In human stereopsis this is accomplished by vergence eye movements. The possibility of using other sources of information to guide eye movement (in particular, texture contours) was mentioned by Grimson [Grim80].

Recently, Grimson has presented results on the interpolation of surfaces over the sparse depth map [Grim81]. He uses a "surface consistency constraint," which states that an absence of zero-crossings implies that the surface shape cannot change radically. Surfaces must then be found which satisfy not only the explicit conditions at the matched feature points, but also the implicit conditions imposed by a lack of zero-crossings between the points. The important assumptions are that the illumination is constant, the albedo is roughly constant, and the surface material is isotropic.

SRI International

(1) Parametric correspondence and chamfer matching: two new techniques for image matching, [Barr77]

A method for matching images to a three-dimensional symbolic reference map is presented. The reference map includes point landmarks, represented with three-dimensional coordinates; linear landmarks, represented as curve fragments with lists of three-dimensional

coordinates; and volumetric structures, represented as wire-frame models. A predicted image is generated from an expected viewpoint by projecting three-dimensional coordinates onto image coordinates and suppressing hidden lines. The predicted image is matched to image features, and the error is used to adjust the viewpoint approximation. The matching is done by "chamfering." The image feature array is first transformed into an array of numbers representing the distance to the nearest feature point and a similarity measure is computed by summing the distance array values at the predicted feature locations.

(2) The SRI road expert: image-to-database correspondence, [Bol178]

The problem of matching an image to a geographic database is studied. The images may vary for several reasons: different camera parameters, lighting conditions, cloud cover, etc. The method that is presented begins with an estimate of the camera parameters, including estimates of uncertainties. It refines the estimated correspondence by locating landmarks in the image and comparing their image locations to their predicted locations. The uncertainties of the camera parameter estimates are modeled as a joint normal distribution. This model implies elliptical uncertainty regions in the image. The location of one feature constrains the uncertainty of others to relative uncertainty regions (These are also ellipses, but are usually significantly smaller than the unconstrained regions). Two kinds of matches between landmarks and image features are used: point-to-point and point-on-a-line. The point-to-point matches yield more information for refining the camera parameters, but the point-on-a-line matches are more numerous and cheaper to find. A modified version of Gennery's calibration method [Genn79] is used to refine the camera parameters.

(3) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, [Fisc81]

A method for fitting a model to experimental data is developed (RANSAC) and applied to the "location determination problem" (i.e., given a set of control points with known positions in some coordinate

frame, determine the spatial location from which an image of the control points was obtained). The method is radically different from conventional methods, such as least-squares minimization, which begin with large amounts of data and then attempt to eliminate invalid points. RANSAC uses a small, randomly chosen set of points and then enlarges this set with consistent data when possible. This strategy avoids a common problem with least squares and similar methods; that is, a few gross errors, or even a single one, can lead to very bad solutions. In practice, RANSAC can be used as a method for selecting and verifying a set of points that can be confidently fit to a model with a conventional method (such a least-squares minimization).

Stanford

(1) Stereo-camera calibration, [Genn79]

A method for determining the relative position and orientation of two cameras from a set of matched points is developed. The calibration accounts for difference in azimuth, elevation, pan, tilt, roll, and focal length. The basic method is a least-squares minimization of the errors of the distances of points in image two from their predicted locations, as determined by their positions in image one and an estimated relative camera model. The nonlinear optimization problem is solved by iterating on a linearization of the problem.

(2) Local context in matching edges for stereo vision, [Arn78]

This approach matches corresponding features instead of matching areas using cross correlation. Two kinds of local feature detectors are used: the Moravec interest operator for sparse points that are used for solving for the relative camera model, and the Heuckel edge operator for a larger number of points that are matched after the relative camera model is known. The approach uses a continuity constraint to resolve ambiguity. If a scene is continouous in three dimensions then adjacent matching edge elements should be continuous in direction and disparity. Intensities on either side of the edge should also be consistent.

The Moravec operator is used to select about 50 points. A coarse/fine search finds matches for some of these points, and Gennery's camera model solver is used to determine the parameters that relate the two camera positions. A dominant plane, which is assumed to be a ground plane, is fit to the matches (few points may lie below the ground plane, some may be above it, and as many as possible lie on it). The Hueckel edge operator is applied to both images (3.19 pixel radius), and the results are transformed into a normalized coordinate system in which points on the dominant plane have zero disparity.

Each edge element in the left picture is matched to nearby candidates in the right image (there are usually about eight candidates) based on the angle and brightness information supplied by the Hueckel operator. Each edge element in the left image is then linked to all its neighbors (in the left image) that seem to arise from the same physical edge. (Two edge elements are neighbors if they are close, have roughly the same angle, and similar brightness. Three or four are typically found.) The linked neighbors of an edge element vote to determine which of the candidate disparities is most consistent (i.e, which is the appropriate matching element in the right image).

Some problems caused by the Hueckel operator are identified (for example, it is unreliable for corners, textured areas, and slow gradients). Relaxation is suggested as a way to use context in a more controlled way (see [Bar79]). The system works well in scenes of man-made objects, but poorly in natural scenes (the opposite of area correlation).

(3) Object detection and measurement using stereo vision, [Genn80]

This study uses stereo or rangefinder data to detect and measure objects, and although it does not deal with the matching problem, it is relevant to the interpolation and interpretation problems. The system is intended for autonomous vehicle guidance and obstacle avoidance.

First, the ground surface is found as described by Arnold in [Arn78]. Above-ground points are clustered with a minimal spanning tree approach, and ellipsoids are fit with a modified least-squares method. Two types of errors are considered: the amount by which the points in a cluster being fit miss lying on the ellipsoid, and the amount by which the ellipsoid occludes any points as seen from the camera. (Orthographic projection, not central projection, is assumed.) In addition, there is an a priori bias to make any small ellipsoids approximately spherical.

After ellipsoids have been fit to the original clusters, it may become apparent that the initial clustering, based on only local information, did not produce a good segmentation. In this case, the initial clusters are either split or merged and another set of ellipsoids is fit to them.

(4) Visual mapping by a robot rover, [Mor79]
Rover visual obstacle avoidance, [Mor81]

This is a study of autonomous vehicle guidance. Severe noise problems are overcome by use of redundancy. An early approach that used only motion stereo was found to be unworkable because of matching errors and uncertain camera models. A subsequent approach used "slider stereo" to obtain nine stereo views. A calibration step determines the camera's focal length and distortion from a digitized test pattern.

An interest operator is used to select good features for matching. First, for each point in the central image it computes the variance between adjacent pixels in four directions over a square (3x3 pixel) neighborhood centered on the point; next, it selects the minimum variance as its interest measure; and finally, it chooses feature points where the interest measure is locally maximal. Intuitively, each chosen point must have relatively high variance in several directions, and must be more "interesting" than its immediate neighbors. The interest operator is used on reduced versions of the images.

A binary search correlator matches 6x6 pixel areas, denoted by features found by the interest operator in the central image, to areas in each side image. The search begins at the lowest resolution (x16 reduction) and proceeds to the higher resolutions. In this way, points chosen from the center view are found in the other eight views. The uncertainty of the depth measurement associated with a match is inversely proportional to the length of the stereo baseline. To obtain more accurate depths, the measurements are averaged by considering each of the stereo baselines obtained from the thirty-six combinations of nine views taken two at a time. A measurement from a particular pair contributes a normal distribution, with a mean at the estimated distance, and a standard deviation inversely proportional to the stereo baseline. The contributions are also normalized according to the correlation coefficients of the matches and according to the degree of y-disparity. (A low correlation coefficient or a large y-disparity causes the peak value of the distribution to be scaled down, thereby reducing the contribution of the depth measurement.) The peak in the sum of these distributions gives a very reliable depth measurement.

Depth measurements are used to help navigate the vehicle, which moves in approximately one-meter increments. Vehicle motion is deduced from depth measurements at two successive positions by comparing the differences of point positions, which should be the same in both views. This approach to navigation is similar to the "bootstrap stereo" method [Han80] described previously.

V CONCLUSIONS

Automated computational stereo cannot simply duplicate the steps and procedures currently employed when a human interpreter is an integral part of the process. There is at present no reasonable way to duplicate the human ability to invoke semantic and physical knowledge to filter out gross errors in the various steps, and especially in the matching steps. Techniques that are highly tolerant of errors (such as RANSAC) will have to be substituted for those that depend on reliable manually filtered data (such as least-squares estimation of camera parameters). Constraints indirectly invoked by the human interpreter must be made explicit and embedded directly into the automated procedures (e.g., the fact that all vertical edges depicted in an image must pass through a common vanishing point). Automated stereo, not limited by the two-image constraint of the human, can partially compensate for the lack of a human knowledge base by "simultaneously" processing a large number of views of a scene to resolve ambiguity, and by approaching some of the problems from a quantitative (model-based) approach rather than the qualitative (constraint-based) approach of humans.

VI ACKNOWLEDGMENTS

The authors thank Oscar Firschein, Bruce Lucas, Takeo Kanade, and William Thompson for their comments and suggestions.

REFERENCES

[Arn78] D. Arnold, "Local context in matching edges for stereo vision," in Proc.: Image Understanding Workshop, May 1978, pp. 65-72.

[Bak80] H. Baker, "Edge-based stereo correlation," Proc: Image Understanding Workshop, April 1980, pp. 168-175.

[Bar79] S. T. Barnard, The Image Correspondence Problem, Ph.D. dissertation, Computer Science Department, University of Minnesota, Minneapolis, Minnesota, 1979.

[Bar80] S. T. Barnard and W. B. Thompson, "Disparity analysis of images," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-2, no. 4, July 1980, pp. 333-340.

[Barr77] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: two new techniques for image matching," Proc.: 5th International Joint Conference on Artificial Intelligence, August 1977, pp. 659-663.

[Bol178] R.C. Bolles, L.H. Quam, M.A. Fischler and H.C. Wolf, "The SRI road expert: image-to-database correspondence," Proc.: Image Understanding Workshop, November 1978, pp. 163-174.

[Duda79] R. O. Duda, D. Nitzan, and P. Barrett, "Use of range and reflectance data to find planar surface regions," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 3, July 1979, pp 259-271.

[Fenn79] C.L. Fennema and W.B. Thompson, "Velocity determination in scenes containing several moving objects," Computer Graphics and Image Processing, vol. 9, April 1979, pp. 301-315.

[Fisc81] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, June 1981, pp. 381-395.

[Gan75] S. Ganaparthy, "Reconstruction of scenes containing polyhedra from stereo pairs of views," Artificial Intelligence Lab., Stanford University, Stanford, California, Memo AIM-272, December 1975.

[Genn79] D. Gennery, "Stereo-camera calibration," Proc.: Image Understanding Workshop, November 1979, pp. 101-107.

[Genn80] D. B. Gennery, "Object detection and measurement using stereo vision," Proc: Image Understanding Workshop, April 80, pp. 161-167.

[Grim79] W. E. L. Grimson and D. Marr, "A computer implementation of a theory of human stereo vision," Proc.: Image Understanding Workshop, April 1979, pp. 41-47.

[Grim80] W. E. L. Grimson, "Aspects of a computational theory of human stereo vision," Proc: Image Understanding Workshop, April 1980, pp. 128-149.

[Grim81] From Images to Surfaces, W. E. L. Grimson, MIT Press, Cambridge, Massachusetts, 1981.

[Guz68] Guzman, A., "Computer recognition of three-dimensional objects in a visual scene," MAC-TR-59 (thesis), Project MAC, M.I.T., Cambridge, MA, 1968.

[Han74] Hannah, M.J., Computer Matching of Areas in Stereo Imagery, Ph.D. dissertation, AIM 239, Computer Science Department, Stanford University, Stanford, California, 1974.

[Han80] M. J. Hannah, "Bootstrap stereo," Proc: Image Understanding Workshop, April 1980, pp. 201-208.

[Hen78] R. L. Henderson, W. J. Miller, and C. B. Grosch, "A flexible approach to digital stereo mapping," Photogrammetric Engineering and Remote Sensing, vol. 44, no. 12, December 1978, pp. 1499-1512.

[Horn75] B.K.P. Horn, "Obtaining shape from shading information," The Psychology of Computer Vision, P.H. Winston, ed., McGraw-Hill, 1975, pp. 115-155.

[Thom66] Manual of Photogrammetry, third edition, M.M. Thompson, ed., American Society of Photogrammetry, Falls Church, Virginia, 1966.

[Limb75] J.O. Limb and J.A. Murphy, "Estimating the velocity of moving images in television signals," Computer Graphics and Image Processing, vol. 4, Dec. 1975, pp. 311-327.

[Luc81] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proc: Image Understanding Workshop, April 1981, pp. 121-130.

[Marr76] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," Science, vol. 194, pp. 283-287, 1976.

[Marr77] D. Marr and T. Poggio, "A theory of human stereo vision," Artificial Intelligence Lab., M.I.T., Cambridge, Massachusetts, Memo 451, November 1977.

[Mor79] H. Moravec, "Visual mapping by a robot rover," Proc. 6th Int. Joint Conf. Artificial Intell., vol. 1, Tokyo, Japan, August 1979, pp. 598-600.

[Mor81] H. Moravec, "Rover visual obstacle avoidance," Proc. 7th Int. Joint Conf. Artificial Intell., vol. 2, Vancouver, Canada, August 1981, pp. 785-790.

[Rob65] L.G. Roberts, "Machine perception of three-dimensional solids," in Optical and Electro-Optical Information Processing, J.T. Tippett et. al., Eds., MIT Press, Cambridge, Massachusetts, 1965.

[Thom66] Manual of Photogrammetry, third edition, M.M. Thompson, ed., American Society of Photogrammetry, Falls Church, Virginia, 1966.

[Wal75] D. Waltz, "Understanding line drawings of scenes with shadows," in The Psychology of Computer Vision, P.H. Winston, Ed., McGraw-Hill, 1975.

[Wit81] A. Witkin, "Recovering intrinsic scene characteristics fro images," SRI Project 1019 Interim Report, Artificial Intelligence Lab., SRI International, 333 Ravenswood Ave., Menlo Park, California, 94025, September, 1981.

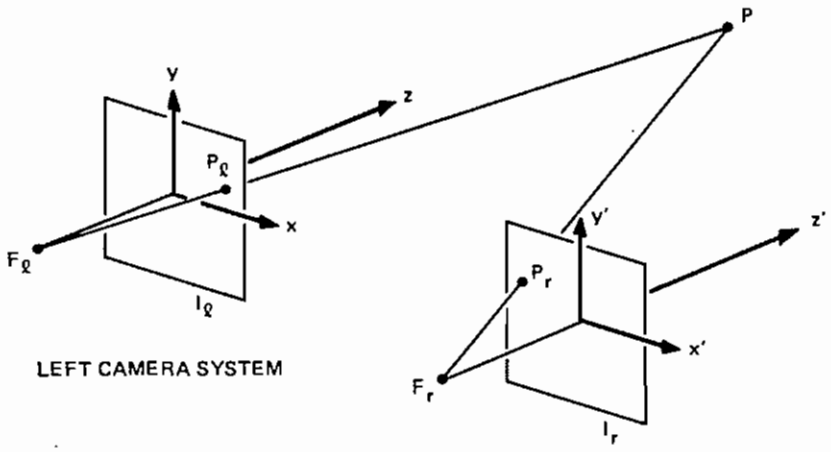

FIGURE 1. STEREO

Two camera systems are shown. The focal points are at F_l and F_r, the image planes are I_l and I_r, and the principal axes are z and z'. A point P in the three-dimensional scene is projected onto P_l in the left image and onto P_r in the right image. The disparity of P is the difference in the positions of its projections onto the two stereo image planes. The disparity of P depends on its location in the scene and on the geometrical relation between the camera systems. In most cases the location of P can be determined from its stereo projections.

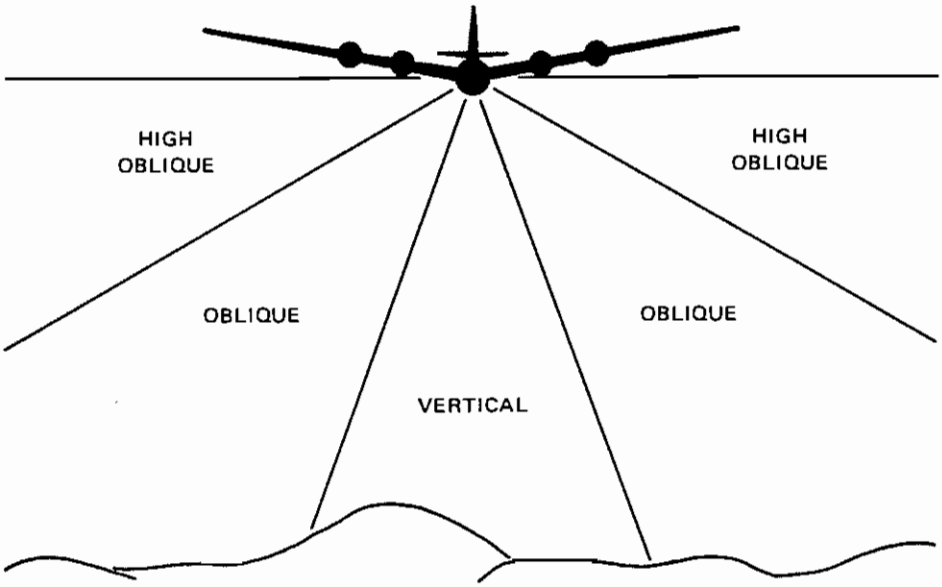

FIGURE 2. VERTICAL AND OBLIQUE AERIAL IMAGERY

Aerial images are usually recorded in long sequences from an aircraft. Vertical images are made with the camera aligned as closely as possible with the true vertical. Oblique images are made by intentionally aligning the camera between the true vertical and horizontal directions. Oblique views that include the horizon are called "high oblique". Even though oblique views are somewhat more difficult to analyze than vertical views, they cover more area and are therefore used for lower cost of image acquisition.

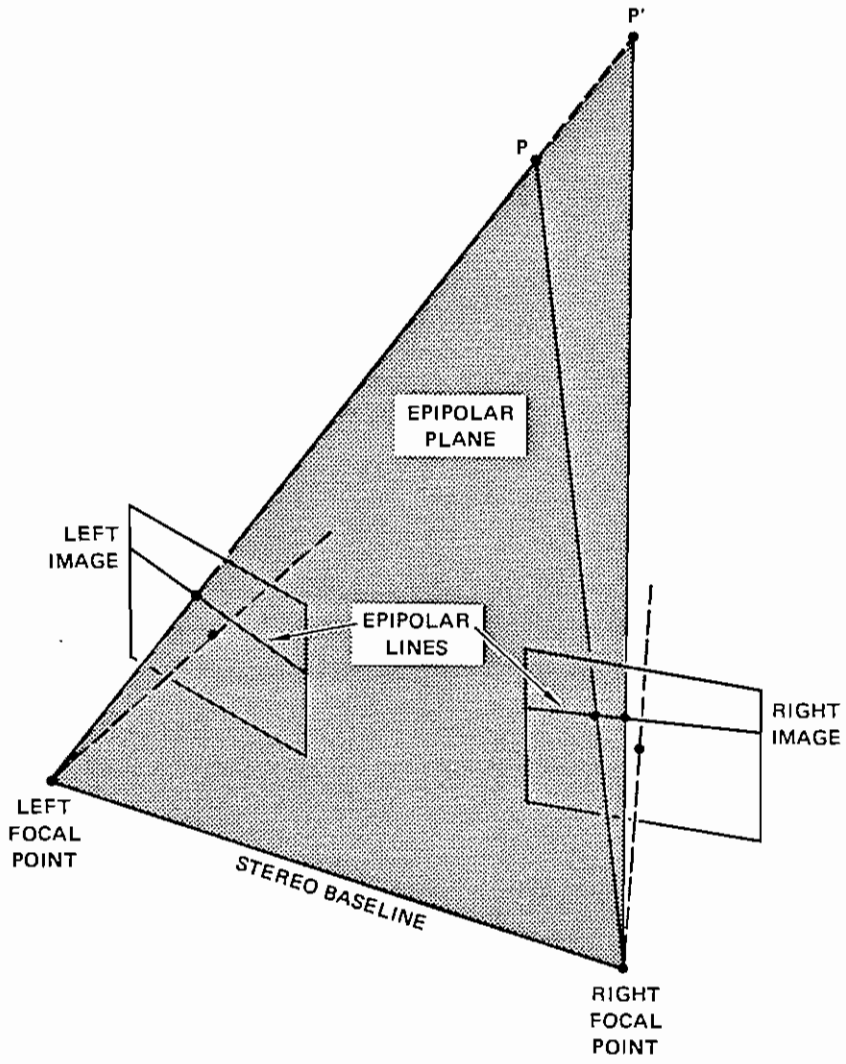

FIGURE 3. THE EPIPOLAR CONSTRAINT

Left and right camera systems are shown. The line connecting the focal points of the camera systems is called the stereo baseline. Any plane containing the stereo baseline is called an epipolar plane. Suppose a point P in the scene is projected onto the left image. Then the line connecting P and the left focal point, together with the stereo baseline, determines a unique epipolar plane. The projection of P in the right image must therefore lie along the line which is the intersection of this epipolar plane with the right image plane. (The intersection of an epipolar plane with an image plane is called an epipolar line.) If the geometrical relationship between the two camera systems is known, we need only search for a match along the epipolar line in the right image.

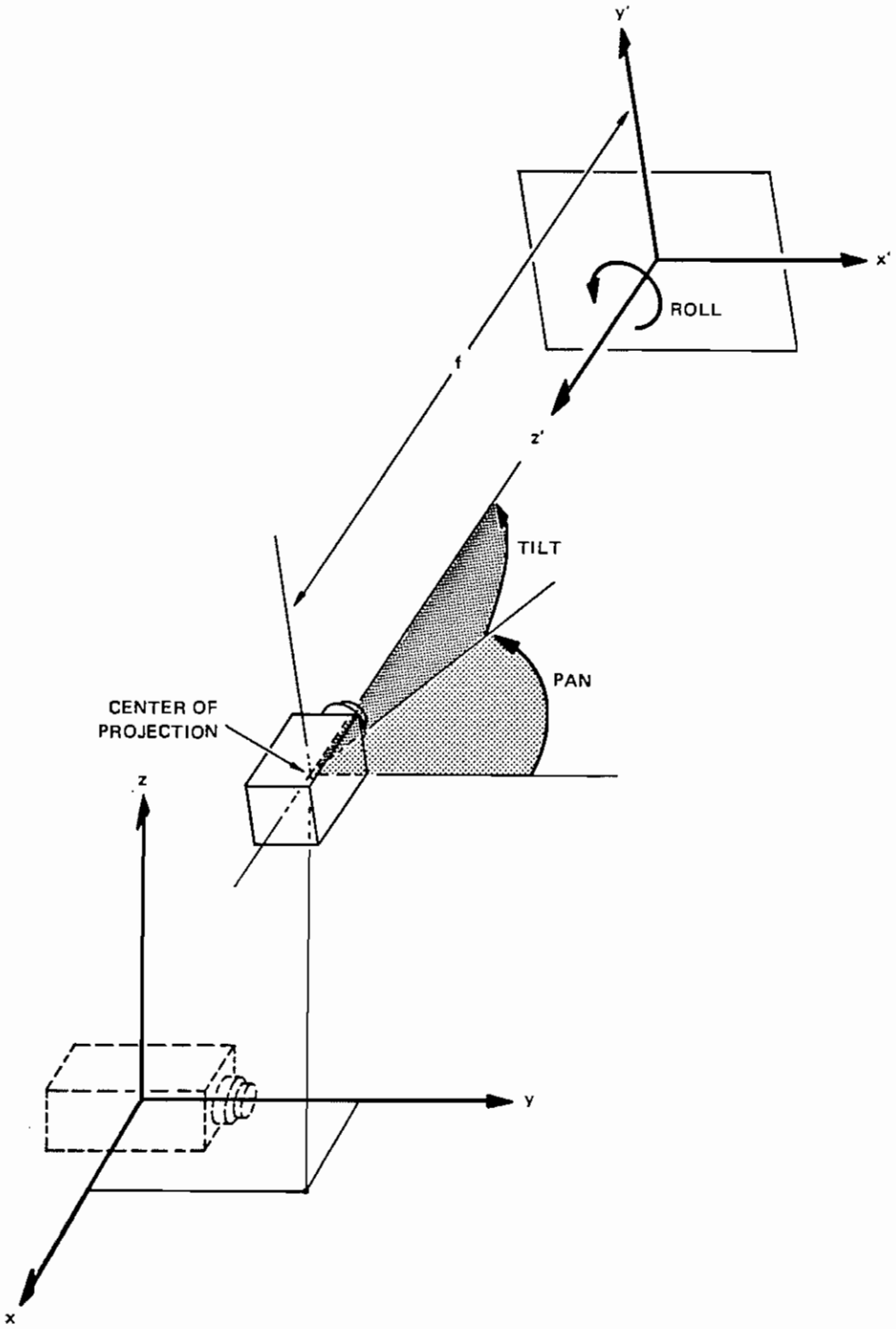

FIGURE 4. CAMERA MODELING

Camera systems are modeled as transforms of three-dimensional coordinate systems. The transforms include translational, rotational, perspective, and scaling components. There are many ways to choose parameters for camera transforms, and this figure illustrates one choice. A reference system is shown with unprimed coordinates. If this reference system is fixed to the scene we have an absolute camera model, and if it is attached to another camera system we have a relative camera model. The camera coordinate system, shown with primed coordinates, is aligned with the image plane. The translational component of the transform is specified by the location of the center of projection (i.e., the focal point of the camera) in the reference system. The rotational component is specified by a pan angle, a tilt angle, and a roll angle. The distance from origin of the camera coordinate system to the center of projection is equal to the focal length f.